DOCUMENT RESUME

ED 330 716 TM 016 302

AUTHOR Garcia-Perez, Miguel A.; Frary, Robert B.
TITLE Item Characteristic Curves: A New Theoretical Approach.
PUB DATE Apr 91
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Decision Making; Equations (Mathematics); Graphs; Guessing (Tests); *Item Response Theory; Knowledge Level; Mathematical Models; Multiple Choice Tests; *Objective Tests; Polynomials; Psychological Characteristics; *Test Use
IDENTIFIERS Finite Element Methods; *Item Characteristic Function; Item Parameters

ABSTRACT

A new approach to the development of the item characteristic curve (ICC), which expresses the functional relationship between the level of performance on a given task and an independent variable that is relevant to the task, is presented. The approach focuses on knowledge states, decision processes, and other circumstances underlying responses to objective tests. Earlier work on finite state models of objective test performance provides the basis for deriving expressions for ICCs that directly account for factors such as examinee willingness to guess, mode of test administration, number of options per item, and response strategy of the examinees. This approach uses a parameterization of ability different from that used in conventional item response theory (IRT) and yields ICCs that are polynomial functions of ability. The degree and coefficients of these polynomials depend in part on certain psychological/circumstantial factors. Examples are provided to demonstrate the means by which differing assumptions about objective test response strategies lead to variation in the shapes of the resulting ICCs. The advantages that IRT could gain from adoption of these ICCs are discussed, and the work that remains to be done before finite state polynomic ICCs can be used in practice is outlined. Some possible extensions to the finite state approach are also discussed. Six figures and a 40-item list of references are included. (Author/TJH)

Item Characteristic Curves: A New Theoretical Approach

Miguel A. García-Pérez
Departamento de Metodología
Facultad de Psicología
Universidad Complutense
Campus de Somosaguas
28023 Madrid
Spain

Tel. (+34)1-394-3061
E-mail PSMET01@EMDUCMS1.EARN

Robert B. Frary (presenter)
Office of Measurement and Research
Division of Information Systems
Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061-0438
U. S. A.

Tel. (+1)703-231-5413
E-mail FRARY@VTVM1.BITNET

Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois

Division D, Session 39.30, April 5, 1991
Marriott Hotel
Los Angeles Room
12:25-1:55

Abstract

A new approach to the development of the item characteristic curve (ICC) is presented, in which knowledge states, decision processes and other circumstances underlying responding to objective tests receive a priori consideration. Earlier work on finite state models of objective test performance provides the basis for deriving expressions for ICCs that directly account for factors such as examinee willingness to guess, mode of test administration, number of options per item, and the response strategy of the examinees. This approach utilizes a parameterization of ability different from that used in conventional item response theory (IRT) and yields ICCs that are polynomial functions of ability. The degree and coefficients of these polynomials depend in part on psychological/circumstantial factors such as those just mentioned or others that may readily be introduced. Examples are provided to show how differing assumptions about objective test responding lead to variation in the shapes of the resulting ICCs. The advantages that IRT could gain from adoption of these ICCs are discussed, and the work that remains to be done before finite state polynomic ICCs can be used in practice is outlined. Some possible extensions to the finite state approach are also discussed.

Item Characteristic Curves: A New Theoretical Approach[1]

Miguel A. García-Pérez
Universidad Complutense
de Madrid

Robert B. Frary
Virginia Polytechnic Institute
and State University

The item characteristic curve (ICC) is a key element of item response theory (IRT). Broadly speaking, an ICC expresses the functional relationship between the level of performance on a given task and an independent variable that is relevant to the task. As applied to ability or achievement testing, where IRT emerged, the ICC expresses the probability of responding correctly to an item as a function of the examinee's (unobservable) ability or knowledge.

Despite being a fundamental feature of IRT models, the true functional form of this relationship must remain unknown. Nevertheless, application of IRT requires the adoption of some mathematical form for the ICC. In Lord and Novick (1968, Section 16.8), some justification is provided for the two-parameter normal ogive, with a derivation of sufficient if rather restrictive conditions for data to be consistent with this model. However, the conditions derived are by no means necessary ones, and Lord (1980, p. 30) stated a preference for considering any particular ICC as representing a basic assumption in its own right, which must be justified empirically. Replacement of the normal ogive with the logistic function was motivated by its ability to mimic the normal ogive while being more tractable mathematically (see Birnbaum, 1968). The further development of the logistic function through the addition of the pseudo-chance parameter, might be said to be theory driven; it was assumed that examinees of very low ability would guess essentially at random on multiple-choice items, resulting in a lower asymptote for the ICC at the probability of a correct guess under these conditions. No further applications of psychological theory have yielded fundamental changes in the mathematical form of ICCs adopted for large-scale IRT applications.

Hambleton and Swaminathan (1985, pp. 9-10) comment on the wide range of IRT models that can be operationalized by simply changing the mathematical form of the ICC but do not mention psychological considerations that might guide these changes. In keeping with Lord's philosophy, they only suggest testing the appropriateness of the choice by conducting goodness-of-fit studies. Samejima (1981, p. 230) pointed out some criteria for choosing among various types of ICCs. However, her main conclusion was that the appropriateness of any of the proposed models depends largely on the guessing behavior of the examinees. McDonald (1982) approached this question more generally, proposing a framework from which many models can be generated by varying the cumulative distribution function to represent the ICC. He also offered no psychological criteria for deciding on the most realistic model, only pointing out the need for statistical tests leading to acceptance or rejection of any particular model.

The foregoing analysis led us to the conclusion that the logistic functions (or any other conventional ICCs for that matter) do not embody psychological theory, in the sense that their a

---

[1]An expanded version of this paper has been accepted for publication in the British Journal of Mathematical and Statistical Psychology under the title, Finite State Polynomic Item Characteristic Curves.

priori appropriateness as ICCs does not follow from a formalization of the processes and variables that are involved in responding to test items. In fact, support for their use only comes a posteriori, once they have been shown to describe data adequately (with the help of suitably estimated parameters). This pragmatic approach to justifying the choice of logistic ICCs was evident in Lord's (1980, p. 31) assertion that "justification of their use is to be sought in the results achieved, not in further rationalizations."

Contrary to prescribing the form of the ICC largely on the basis of expediency, we adopt here an explanatory approach to the generation of mathematical expressions that can be used as ICCs. Adopting a different parameterization of examinee ability and item difficulty, and starting from (replaceable) assumptions about examinee behavior and item characteristics, finite state theory allows the derivation of expressions for the probability of responding correctly to a test item. This approach produces ICCs that are primarily dependent on ability and difficulty, just as is the case for conventional ICCs. However, certain aspects of their mathematical expression also depend on other variables. These variables, not incorporated into conventional IRT models, include the number of options per item, examinee willingness to guess when uncertain, the response strategy followed by the examinees, the format of administration of the test, and potentially other item characteristics. This point is illustrated by providing ICCs for different situations. All of these ICCs are polynomial functions of ability, and we refer to IRT models built around them as *finite state polynomic models*.

The goal in this paper is to introduce these new ICCs and to compare them with logistic ICCs from a number of points of view, with special attention to the different parameterizations underlying each type of ICC and their theoretical foundations.

Finite State Theory and Finite State Polynomic ICCs

The assumptions and definitions underlying finite state modelling of objective test performance have been thoroughly dealt with elsewhere (Garcfa-Pérez, 1987, 1989a, 1989b, 1990; Garcfa-Pérez & Frary, 1989). To avoid repetition, only a brief account of the theory will be supplied here, which will suffice for the development to follow. However, a reading of Garcfa-Pérez (1987) and Garcfa-Pérez and Frary (1989) will provide a more detailed justification for some of the assumptions and definitions that will now be introduced.

The term "statement" is central to the finite state approach. In the context of multiple-choice testing, a statement is any sentence resulting from adding to the item stem one of its available options. Finite state theory defines the level of knowledge of an examinee as the proportion of statements about a subject matter whose truth value he/she knows. This characterization of knowledge is akin to Falmagne and Doignon's (1988) definition of a *knowledge state* with respect to a body of information. In addition, the theory assumes that, when facing a multiple-choice item, the examinee makes *independent* attempts to classify every single available option as true or false. This process gives rise to a finite set of knowledge states about the item, ranging from total ignorance through several degrees of partial knowledge to total

knowledge. These states, in conjunction with the guessing strategy that the examinee adopts, determine whether a conventional multiple-choice item will be answered correctly, answered incorrectly, or left unanswered.

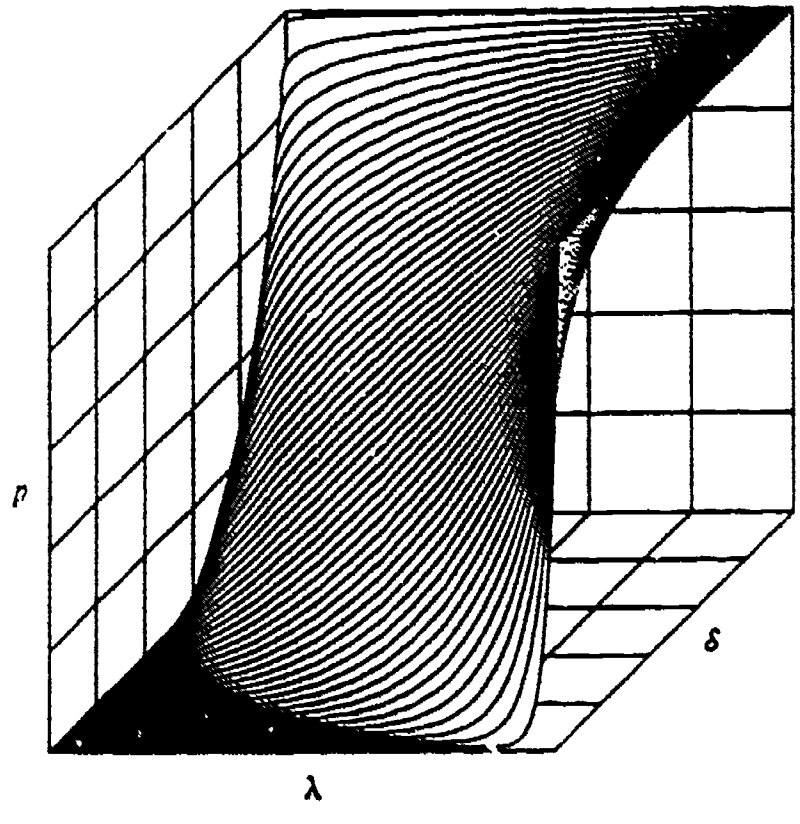
When the process just described is translated into mathematical terms, expressions for the probability of these observable response outcomes can be derived. A two-stage modelling process accomplishes this goal. First, an expression is adopted for the probability that the truth value of a statement will be known. Then, expressions are derived for the probability that any given response outcome will occur. These two stages will now be dealt with separately.

Probability of Knowing the Truth Value of a Statement

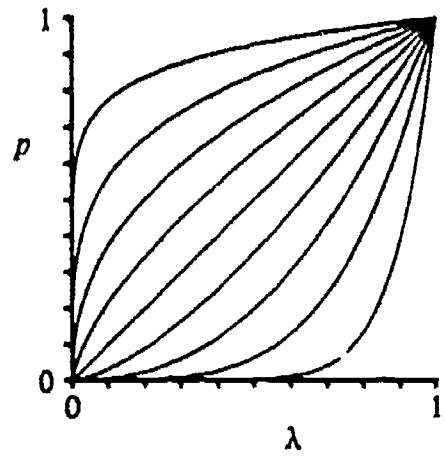
Let λ ($0 \leq \lambda \leq 1$) be the true proportion of statements about the subject matter whose truth value an examinee knows. This is the (unidimensional) latent variable representing the examinee's level of knowledge or ability, and it is also the probability that he/she will know the truth value of a *randomly drawn* statement in a multiple-choice item. We call it λ instead of the usual θ in IRT for the sake of consistency with our previous work and to stress the fact that it is not in some sense interchangeable with θ. Not only do λ and θ span different ranges, they also have different relationships to performance, as will be seen. While λ is *not* the probability of answering an *item* correctly, it is clearly related to this probability, as will be shown. Also influencing the probability of answering correctly is the presence of topics in the subject matter of interest, represented by items on the test, that are easier or more difficult than others. Therefore, it would seem an oversimplification to assume that an examinee has a probability λ of knowing whether an option is true or false as applied to a given item stem, not taking into consideration the difficulty of the question being asked. So let δ ($0 < \delta < 1$) represent item difficulty with values closer to 1 the easier the item, as is the case for the classical difficulty parameter. We call it δ instead of b, as is usual in logistic ICCs, because this parameter does not have the same meaning nor the same effect as b, as will be seen below. To take item difficulty into account, we let the probability, p, that an examinee with ability λ knows the truth value of an option in an item of difficulty δ be

$$p = \lambda^{1/\delta - 1}. \qquad (1)$$

Figure 1a shows a three-dimensional plot of this power function of the inverse of item difficulty, and Figure 1b shows sections of this function for items of selected difficulties. Note that, for any given λ, p increases with decreasing item difficulty (increasing δ). Note also that $p > \lambda$ when $\delta > .5$ and $p < \lambda$ if $\delta < .5$ while $\delta = .5$ makes $p = \lambda$. Our choice of the functional relationship of Equation 1 was limited to functions such that, as δ increases from 0 to 1 (though not taking on these extreme values), p increases gradually and monotonically from a value of 0, attaining the value λ when $\delta = .5$, and the value 1 when $\delta = 1$. It is similar to the power functions that appear in psychophysics (Atkinson, 1982) and is consistent with attempts to establish links between test theory and psychophysics (Mosier, 1940, 1941; Hutchinson, 1977).



(a)



(b)

Figure 1. (a) 3-D plot of Equation 1. The origin of coordinates is at the lower left corner of the rhomboid over which the surface ascends. Examinee ability (λ) increases from 0 to 1 along the horizontal axis. Item difficulty (δ) increases from 0 to 1 along the 45° tilted axis. The height of each point in the surface is the probability of correct classification of an option in an item of difficulty δ by an examinee of ability λ as given by the plane coordinates of its vertical projection. Reference grid lines are spaced at intervals of .2 units.

(b) 1-D cuts of the probability surface at several item difficulty values. From top to bottom, the curves represent the probability of correctly classifying an option within an item as a function of ability for items of difficulties $\delta = .9, .8, .7, .6, .5, .4, .3, .2$, and .1.

It may be noted that Equation 1 is actually the ICC for a very simple type of item to which examinees respond under very restricted conditions, namely, true-false items at which they never guess. As such, this function was chosen arbitrarily (but in conformity with the criteria just outlined). If true-false responses with omissions in the absence of knowledge could exist in reality, there is then the question of how well they would fit the ICC of Equation 1. It is possible that they would fit some other function better, for example, a bilinear function, such as those used in Frary (1985) and in García-Pérez and Frary (1989). We leave open the question of the appropriateness of Equation 1 but will adopt it for further development in this paper, because it is plausible, mathematically tractable, and not at all critical to our main argument. Any other function meeting the criteria outlined above could be used and would be preferable if it led to a better fit of real data. What Equation 1 (or a substitute for it) represents is a prototypical ICC. We will show how it may be used as a "building block" in the production of ICCs accounting for various sets of circumstances associated with multiple-choice testing, circumstances that go beyond the simple case for which Equation 1 might be appropriate.

Probability of Each Response Category to a Multiple-Choice Item

To demonstrate that application of Equation 1, we will assume a set of conditions associated with a multiple-choice test. These assumptions were chosen to specify a rather comprehensive set of testing circumstances. However, to facilitate preliminary development, the assumptions are basically simple. (As a result, some of them may not seem highly plausible, though they are by no means impossible.) Following development of the ICC for these preliminary assumptions, various ones will be modified in turn in the next section, and the resulting ICCs will be derived. The preliminary assumptions are as follows:

i local independence across items.

ii independence of options. This means that options within an item must be independently classifiable by examinees as if they actually were independent true-false items. Thus, correct classification of fewer than all of the distractors must not lead the examinee to infer what the correct answer is if he/she does not know it. Unavoidably, of course, correct classification of the answer must lead to classification of previously unclassified options as distractors, but this situation is handled appropriately by our procedure, as will be seen.

iii test with three-option items.

iv distractors that are equally attractive as the correct answer to each item. This means that, for a given examinee, the probability of being able to classify a randomly chosen correct option is the same as for a randomly chosen distractor.

v examinee behavior such that the occurrence of (random) guessing among unclassified options is determined only by each examinee's individual (overall) willingness to do so irrespective of the number of distractors that he/she has identified on a particular item. Individual differences exist in willingness to guess at random, so let γ ($0 \leq \gamma \leq 1$) repre-

sent this willingness as the probability that a (specific) examinee will guess at random when the correct answer is not known.

vi conventional administration of the test, i.e., asking examinees to mark the alternative believed to be correct for each item, but without advice as to how the test will be scored or regarding the guessing strategy required for score optimization. (This lack of information would be consistent with the guessing behavior assumed in Assumption v.)

As a consequence of all of the above assumptions, an examinee of ability λ has a probability p of knowing whether each of the options of an item of difficulty δ is true or false. Thus, he/she will be able to classify the answer and some number of distractors for the item depending on λ and δ, as is clear from Equation 1. But it may happen that this knowledge will be insufficient to permit marking the correct answer with assurance. In this case, the examinee is free to guess at random among the unclassified options.

With these considerations in mind, our task is to derive a mathematical expression for the probability of getting the correct answer to such an item as a function of λ, γ, and δ. Use of a tree diagram to describe the possible sequences of events when responding to that multiple-choice item facilitates this task. The tree diagram for three-option items responded to under the directions in Assumption vi has been presented and described in detail in Garcfa-Pérez and Frary (1989, Figure 1) for the special case in which the probability of being able to classify each option is simply λ. Figure 2 is an adaptation, in which we have replaced λ with p in accordance with the development above. Note that assumptions ii-vi have been taken into account in constructing this diagram. (Local independence across items is only needed to collapse data from all items in the test.) Independence of options is necessary in order that the links of the tree diagram signifying classification of options within a path be statistically independent. The three options in the item give rise to the eight-branch structure that is represented by the first three links. Four possible states of knowledge regarding the item arise from this branching: correct classification of all three options (total knowledge), correct classification of two options only (high partial knowledge), correct classification of a single option (low partial knowledge), and correct classification of no option (total ignorance). In case of total knowledge, the examinee always gives the correct answer to the item. In case of partial knowledge, Assumption iv leads us to apply a probability of k/n that the correct answer to an n-option item is among k ($0<k<n$-1) classified options. If exactly n-1 options are classified, then the correct answer is always given since it is either included among the classified options or it is the only option that is not identified as a distractor, which means that it is the answer (disregarding the possibility of misinformation). If the correct answer is not known in other cases of partial knowledge or in the case of total ignorance, the examinee, according to Assumption v, may either guess at random (succeeding or failing) or leave the item unanswered. Finally, Assumption vi results in these states of knowledge leading to three possible response outcomes: correctly answered item, wrongly answered item, or unanswered item, as shown to the right of each path by C, W, and U, respectively. Also shown to the right of each path is the probability of that particular sequence of events, which is the product of the probabilities of each link within that path. Since there are several sequences that result

in the same outcome, the sum of all probabilities of paths with the same result is the actual probability of that outcome. Then, we get

$$c = p^3 + 3p^2(1-p) + p(1-p)^2 + p(1-p)^2\gamma + (1-p)^3\gamma/3, \quad (2a)$$
$$w = p(1-p)^2\gamma + 2(1-p)^3\gamma/3, \quad (2b)$$
$$u = 2p(1-p)^2(1-\gamma) + (1-p)^3(1-\gamma), \quad (2c)$$

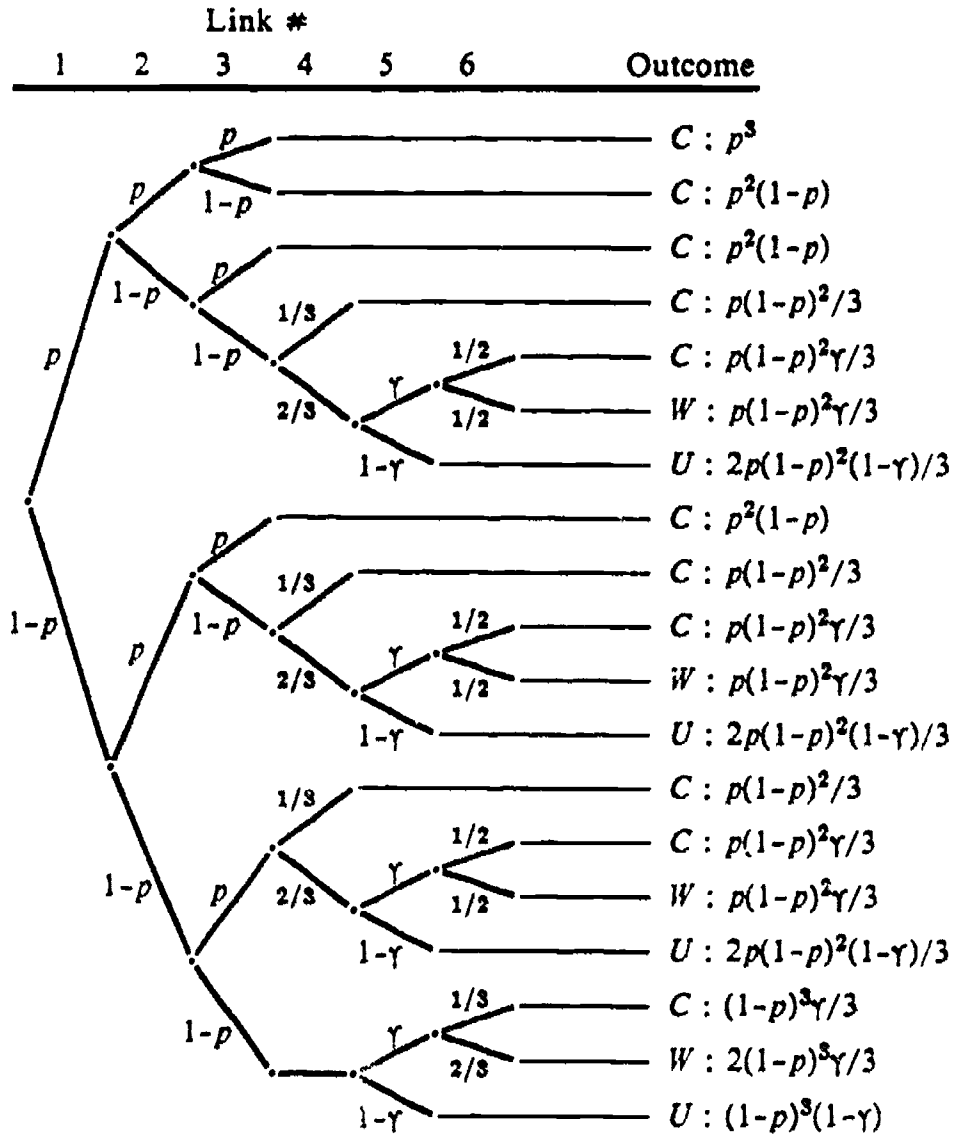


Figure 2. Tree diagram describing the possible sequences of events when responding to a test item when Assumptions i-vi in the text hold. The first three links represent attempts at classifying each option as right or wrong. In the paths where they appear, the fourth links represent inclusion of the correct answer among the classified options, the fifth links represent decisions as to guessing, and the sixth links represent the outcomes of those guesses. Each path results in a response outcome that is represented to the right of the path by either *C*, *W*, or *U*. Also shown to the right of these letters is the probability of the sequence in each path.

in which c, w, and u denote the probabilities of the response outcomes designated by the corresponding uppercase letters. Note that Equation 2a expresses the probability of responding correctly to an item as an explicit function of ability and item difficulty (and, also, guessing propensity). Therefore, like Equation 1, it is an ICC. Note also that Equation 2a embodies Assumptions ii-v, since they were used in the construction of the tree diagram from which this equation comes. These assumptions cover the test administration format (Assumption v), the number of options per item (Assumption iii), the guessing behavior of examinees (Assumption v), and other item characteristics (Assumptions ii, and iv). To illustrate the appear-

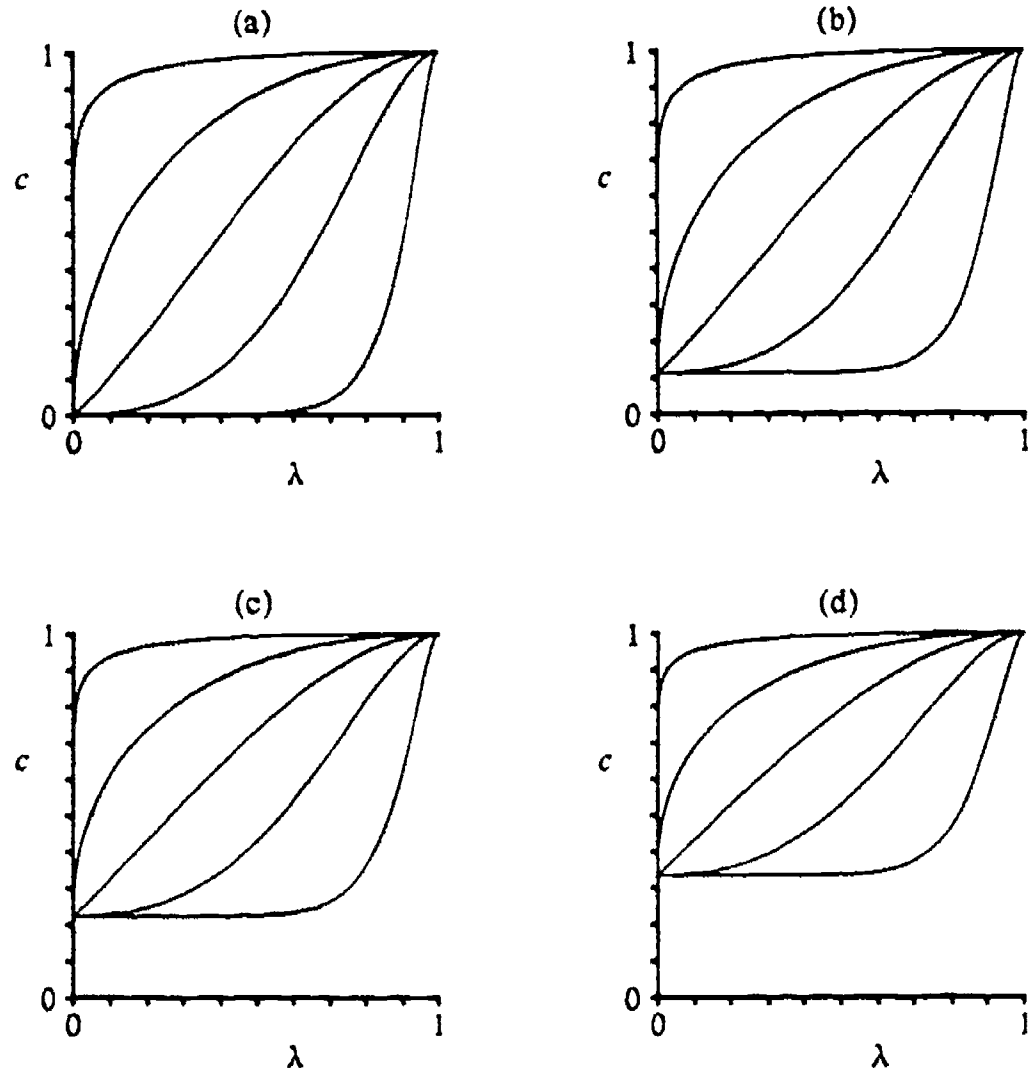


Figure 3. Finite state polynomic ICC given by Equation 2a. In each plot, curves represent, from top to bottom, ICCs for items with $\delta = .9$, .7, .5, .3, and .1. (a) $\gamma = 0$. (b) $\gamma = .33$. (c) $\gamma = .67$. (d) $\gamma = 1$.

ance of this ICC, Figure 3 shows plots of c as a function of λ for items of various difficulties and examinees of differing guessing propensities. Note that the probability of a correct answer to the item increases with increasing guessing propensity and increases too with decreasing

item difficulty (increasing δ). Interestingly, not only does this procedure produce ICCs tailored to the test format and the testing situation under consideration; it also gives rise to other functions relating ability (λ) to the probability of a wrong response or the probability of leaving the item unanswered. These functions could have important theoretical implications for polychotomous response models. In the next section, we further illustrate the flexibility of finite state modelling to derive proper ICCs by considering items for which different variations on Assumptions i-vi hold.

Finite State Polynomic ICCs for other Situations

While not aiming to produce an atlas of polynomic ICCs, we explore here the consequences of varying the assumptions that led to Equations 2a-2c in the previous section. Our main goal is to show how various assumptions representing characteristics of the testing situation can be incorporated into this procedure to derive matching ICCs. In order to make clear what these effects are, we will modify each assumption in turn and produce corresponding ICCs for the new situations. However, we will skip Assumptions i and ii. Local independence across items is retained because, as noted earlier, it is required for collapsing data across all items in the test. Assumption ii concerning independence of options might be removed. As mentioned above, this assumption implies that correct classification of fewer than all of the distractors must not lead to correct classification of the answer when it is not known. Violation of this assumption could easily be handled by the model, but we will not consider this case here because test items with this characteristic would be considered logically defective both by examinees and by score users.

Assumption iii: Number of Options per Item

Let us assume that there are four rather than three options per item but that the other assumptions listed above remain the same. This change has one consequence in the tree diagram, namely, that there are four rather than three links corresponding to options. The resulting tree diagram has 48 paths instead of the 19 paths of the diagram in Figure 2. We do not show it here due to its complexity, but, nevertheless, upon constructing it, it can easily be seen that

$$c = p^4 + 4p^3(1-p) + 3p^2(1-p)^2 + 3p^2(1-p)^2\gamma/2 + p(1-p)^3 + p(1-p)^3\gamma + (1-p)^4\gamma/4, \quad (3a)$$
$$w = 3p^2(1-p)^2\gamma/2 + 2p(1-p)^3\gamma + 3(1-p)^4\gamma/4, \quad (3b)$$
$$u = 3p^2(1-p)^2(1-\gamma) + 3p(1-p)^3(1-\gamma) + (1-p)^4(1-\gamma). \quad (3c)$$

Figure 4 shows plots of Equation 3a for the same values of γ and δ as in the corresponding plots for Equation 2a in Figure 3. Comparing Figures 3 and 4, it can be readily seen that any curve for 4-option items is always below the corresponding curve for 3-option items. That is, other things being equal, increasing the number of options in the item has the effect of lowering the probability of an examinee's correctly responding to it.

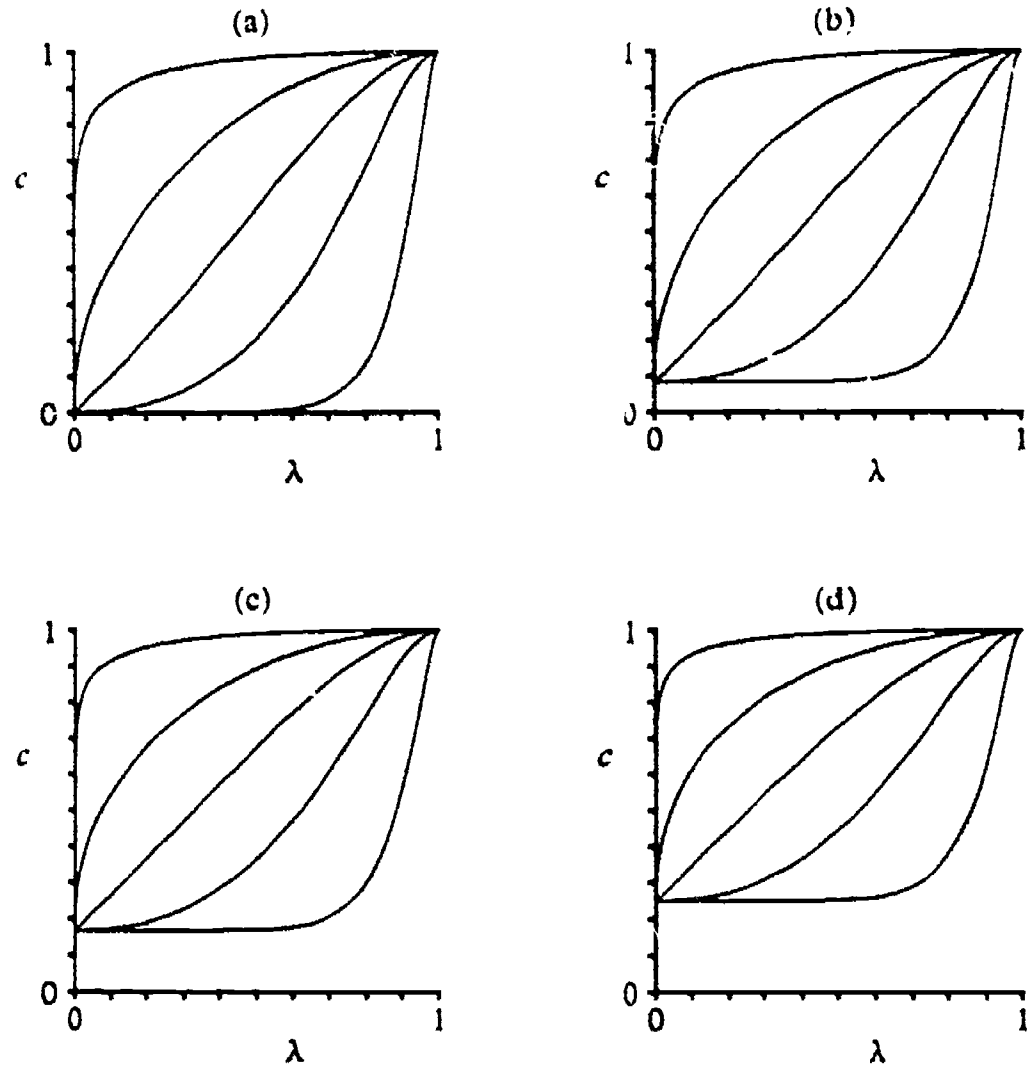


Figure 4. Finite state polynomic ICC given by Equation 3a. In each plot, curves represent, from top to bottom, ICCs for items with $\delta = .9, .7, .5, .3,$ and $.1$. (a) $\gamma = 0$. (b) $\gamma = .33$. (c) $\gamma = .67$. (d) $\gamma = 1$.

Assumption iv: Identifiability of Distractors

Items are sometimes found one of whose distractors is much more readily classifiable than the other options. Finite state modelling can easily accomodate items of this sort by assuming that if a single option is classified, then it is a distractor. In the tree diagram of Figure 2, this assumption means that the probability that a single classified option is the correct answer is zero, while the probability of the correct answer being among k $(1 < k < n\text{-}1)$ classified options remains k/n. When three-option items are considered (i.e., for $n = 3$), only the first part of this statement applies (since there is no integer k such that $1 < k < 2$), and it results in branches with probability ⅓ at the fourth link in Figure 2 being changed to 0 and branches with

probability ⅔, also at the fourth link, being changed to 1. After these modifications, including the deletion of paths one of whose links has been assigned a probability of zero, we get

$$c = p^3 + 3p^2(1-p) + 3p(1-p)^2\gamma/2 + (1-p)^3\gamma/3, \quad (4a)$$
$$w = 3p(1-p)^2\gamma/2 + 2(1-p)^3\gamma/3, \quad (4b)$$
$$u = 3p(1-p)^2(1-\gamma) + (1-p)^3(1-\gamma). \quad (4c)$$

Figure 5 shows plots of Equation 4a for the same values of γ and δ as above. There are a number of different assumptions regarding the identifiability of distractors that can be used in place of either of the two we have considered here. An example of another can be found in García-Pérez (1990).

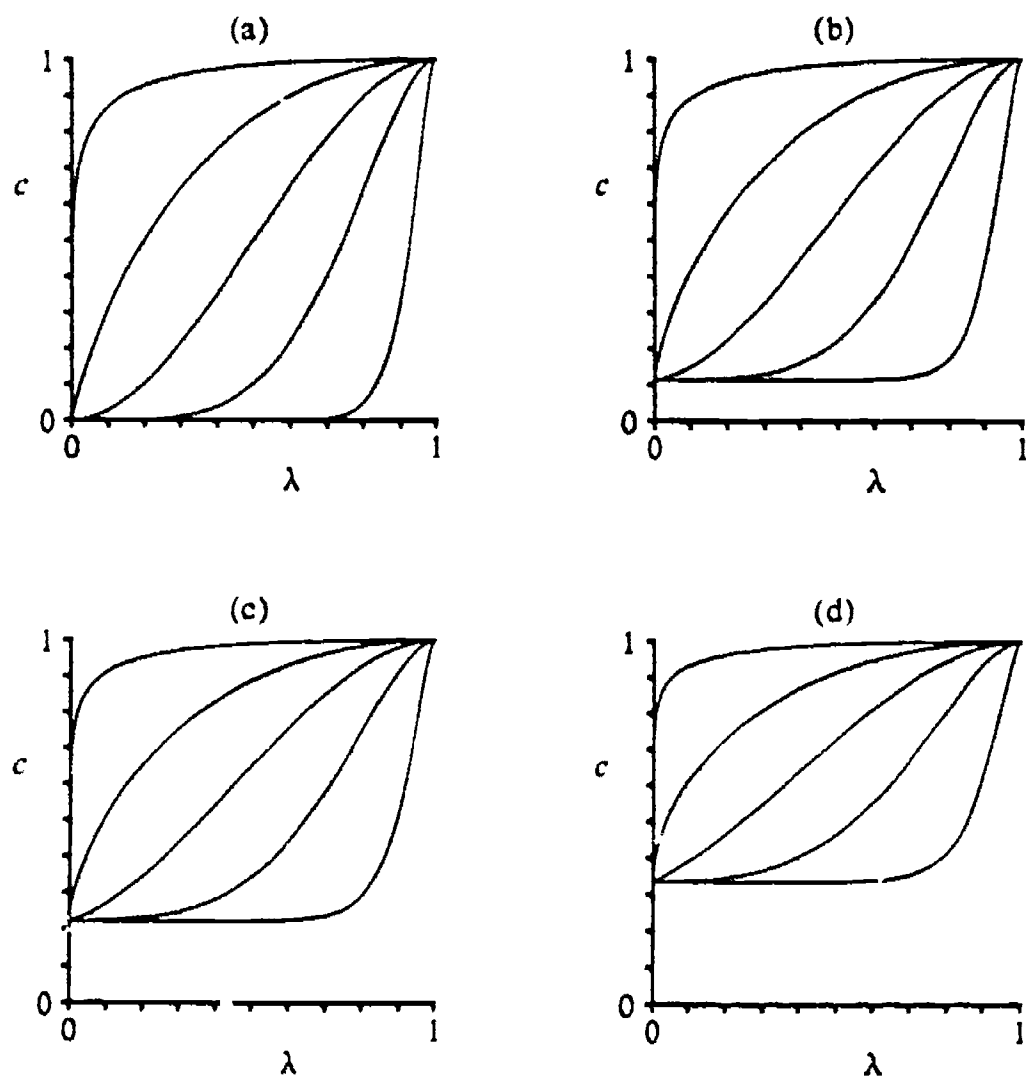


Figure 5. Finite state polynomic ICC given by Equation 4a. In each plot, curves represent, from top to bottom, ICCs for items with $\delta = .9, .7, .5, .3$, and .1. (a) $\gamma = 0$. (b) $\gamma = .33$. (c) $\gamma = .67$. (d) $\gamma = 1$.

Assumption v: Response Behavior

Now we will assume that the examinees take the test under directions to answer every item regardless of knowledge. This practice will serve to eliminate a construct-irrelevant contaminant, namely, γ. If the examinees comply with these directions, all of them will behave as if $\gamma=1$ regardless of anyone's particular willingness to guess. Under these circumstances, the tree diagram is a simplified version of that in Figure 2 with $\gamma=1$ everywhere. There, only two possible response outcomes remain, whose associated probabilities can be shown to be

$$c = p^3 + 3p^2(1-p) + 2p(1-p)^2 + (1-p)^3/3, \quad (5a)$$
$$w = p(1-p)^2 + 2(1-p)^3/3. \quad (5b)$$

Equations 5a and 5b are the same as Equations 2a and 2b when $\gamma=1$. (Also, $\gamma=1$ makes $u=0$ in Equation 2c.) Therefore, the ICCs described by Equation 5a for selected values of δ are the same as those arising from Equation 2a for examinees with $\gamma=1$ (see Figure 3d). Note that the ICC represented in Equation 5a is the only one throughout this paper that applies to a dichotomous response model. As another example of how different response behaviors can be modelled using finite state methods, response behavior appropriate for formula scoring has been considered in García-Pérez and Frary (1989).

Assumption vi: Format of Administration

When the administration format varies, the main structure of the tree diagrams remains basically the same, since it represents knowledge and behavior that are largely independent of the administration format. The only difference is in the assignment of paths to the response categories that are possible under the particular format under consideration. We will illustrate this point by considering a test administered under answer-until-correct (AUC) directions. In this case, examinees continue selecting options until the correct answer is chosen (see Hanna, 1975). Figure 6 shows the tree diagram for this situation. It differs from the diagram in Figure 2 in that the guessing link has been omitted, since examinees behave as if $\gamma=1$. Also, paths formerly leading to correct responses continue to do so as correct responses at the first attempt (C_1). Some of those formerly leading to wrong responses now result in correct responses at the second attempt (C_2), and some others result in correct responses at the third attempt (C_3). Finally, all formerly unanswered items result now in C_1, C_2, or C_3 (a correct response on the first, second or third attempt). Therefore, from Figure 6,

$$c_1 = p^3 + 3p^2(1-p) + 2p(1-p)^2 + (1-p)^3/3, \quad (6a)$$
$$c_2 = p(1-p)^2 + (1-p)^3/3, \quad (6b)$$
$$c_3 = (1-p)^3/3, \quad (6c)$$

in which c_1, c_2, and c_3 denote the probabilities of the response outcomes designated by the corresponding uppercase letters. Note that Equation 6a is the same as Equation 5a. This is because examinees make their first attempt under AUC directions under the same circum-

stances as under conventional directions, with answer-every-item behavior, regardless of the number of options per item. The curves in Figure 3d apply in this case too. However, Equations 6b and 6c are also relevant, providing relationships between ability and probability of a correct response on the second and third attempts. Thus, AUC directions give rise to polychotomous response models, while answer-every-item behavior under conventional administration of the test results in a dichotomous response model. This fact has a substantial bearing on parameter estimation; the outcome of a second attempt at answering a three-option item provides information beyond that which is available when the test is administered under answer-every-item directions. And, clearly, the number of additional sources of information that can be used in parameter estimation increases with the number of options per item when the AUC format of administration is considered instead of the conventional one.

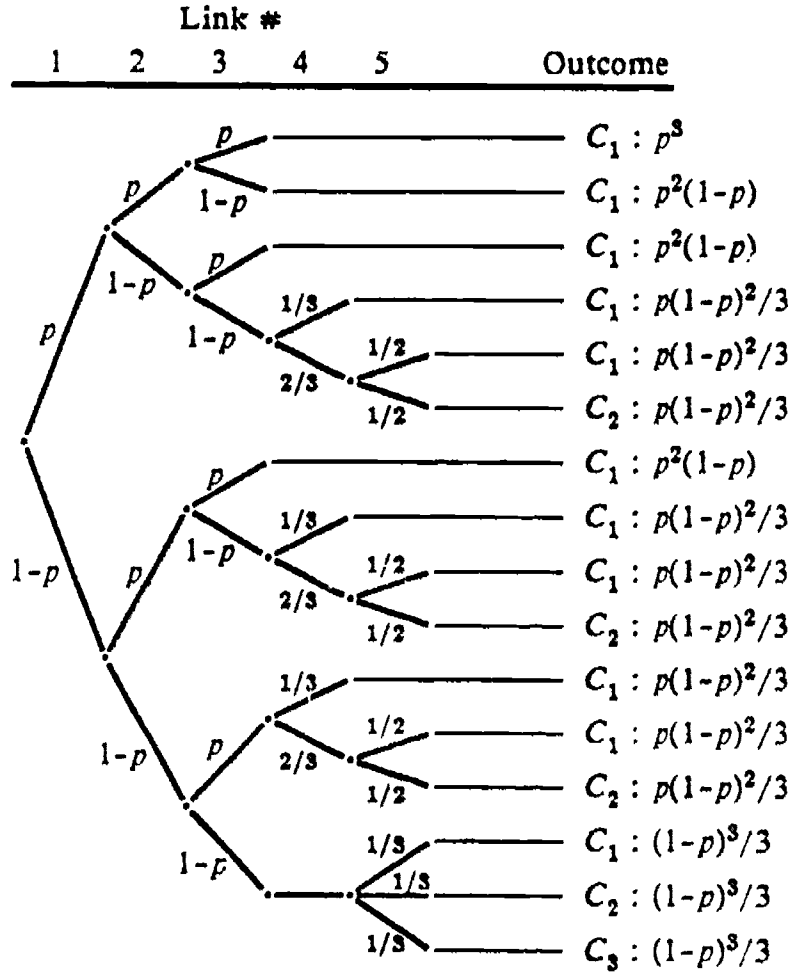


Figure 6. Tree diagram for the same situation as in Figure 2, except that the test is now responded to under AUC directions. The only differences between this diagram and that in Figure 2 are that the guessing link has been removed here since $\gamma = 1$ and that responding with total ignorance may result in one among three response outcomes. Note, however, that the response outcomes are different from those in Figure 2.

Comparison with Conventional ICCs

As shown in the foregoing development, finite state polynomic ICCs follow naturally from a mathematical description of objective test performance that starts from a parameterization different from that implicit in logistic ICCs. As a consequence, the number of parameters and their meanings differ substantially in logistic ICCs in comparison with those arising from finite state modelling. Although the two types of ICC should eventually be compared empirically, this section is devoted to a theoretical analysis of the differences between the item and examinee parameters of each type of ICC.

Polynomic δ versus Logistic b and a

The difficulty parameter in logistic ICCs is the point on the ability scale at which an examinee has a probability of answering the item correctly that is half-way between the lower asymptote and 1. It is clear, however, that δ does not have this interpretation, as revealed by inspection of the plots in Figures 3-5. Further, variation in the logistic difficulty parameter merely produces a horizontal displacement of the ICCs along the ability scale, while variation of δ in finite state polynomic ICCs also varies the steepness of the curves (see Figures 3-5). Therefore, δ in finite state polynomic ICCs accomplishes the same effects as both b and a together in logistic ICCs.

This characteristic of δ can best be appreciated if one considers the different meanings assigned to item difficulty and discrimination parameters in classical test theory as opposed to IRT. In classical theory, item difficulty is defined as the ratio of the number of examinees who answer the item correctly to the number of examinees who attempt it, and item discrimination is related to the number of distinctions that can be made among examinees based on responses to the item. Obviously, these two parameters are not independent, and there is a well-known inverted-U-shaped relationship between them: the number of distinctions that can be made increases as item difficulty approaches a medium value, and it rapidly decreases as difficulty approaches either of its extreme values. Although very easy or very difficult items have low discriminating power from this point of view, it is also true that a fairly easy (alternatively, difficult) item, which will not serve to distinguish among examinees of high (alternatively, low) ability, can nonetheless be useful to distinguish among examinees of low (alternatively, high) ability. The classical item discrimination parameter is deficient in this sense, since it fails to capture the fact that items may have the same discrimination capability, but at different ability levels. Conventional IRT attempted to remedy this situation by adopting different definitions for difficulty and discrimination that made these parameters independent of each other. The difficulty of an item was redefined as the ability needed to have a 50% chance of answering that item correctly, and item discriminating power was redefined to reflect the accuracy with which two examinees can be distinguished when they have abilities slightly above and below the ability designated as item difficulty. These two parameters are known to be related to the point of inflection and to the slope at this point on a conventional ICC, and either of these two values can be varied independently of the other.

Unlike the way either classical theory or IRT handles difficulty and discrimination, finite state modelling implicitly assumes that the ability of an item to distinguish among examinees (i.e., its discriminating power) is a consequence of the *interaction* between the difficulty of the item and the ability of the examinees who respond to it. As can be seen in Figures 3-5, the inflection points of finite state polynomic ICCs move to higher positions on the ability scale as item difficulty increases (δ decreases). Thus, for these ICCs, items discriminate (in an IRT sense) at ability levels that are directly related to their difficulty. At the same time, the slopes of the inflection points are such that the inverted-U-shaped relationship between the classical difficulty and discrimination indices will hold.

Polynomic γ versus Logistic c

As Thissen and Steinberg (1986, Equation 5) show, the third parameter in logistic ICCs results from assuming that the probability of guessing correctly is a fraction of the probability of really not knowing the answer. This fraction is nominally regarded as an item characteristic, with, for very low ability examinees, a maximum equal to the inverse of the number of options. However, lower values are often assigned to obtain better fits of the data. The need for this lowering in an item parameter has been assumed to reflect an *examinee* characteristic, namely, being misinformed or gullible with respect to certain distractors. Departures from the inverse of the number of options have also been attributed to differences in guessing propensity on the part of the examinees by Mislevy and Bock (1982, pp. 727-728) who wrote that, owing to these differences, "the Birnbaum three-parameter model for dichotomous items, which posits for each item a guessing probability that is constant over all examinees, will be in error." As they point out immediately afterwards, application of that model tends to over-reward frequent guessers and under-reward examinees who tend to refrain from guessing. As far as a comparison with the parameterization underlying finite state polynomic ICCs is concerned, the important point is that the third parameter of logistic ICCs is forced to contain *both* item and examinee components, despite being regarded nominally as only an item parameter.

Finite state modelling treats these two influences separately. Willingness to guess is incorporated as a second examinee parameter, γ, which converts the polynomic ICC into an *item characteristic surface*. The polynomic ICC lower asymptotes are then determined for every particular γ, with the lower asymptote reaching its minimum at 0 when $\gamma=0$ and reaching its maximum at the inverse of the number of options (see Figures 3-5) when $\gamma=1$. Actually, it makes little sense to speak about lower asymptotes at dimensional cuts of two-dimensional functions, although any cross-sectional profile of the item characteristic surface at a selected γ will render a true ICC, and it is helpful to keep this in mind.

Because finite state polynomic ICCs have γ as a second examinee parameter, the effects of variations in guessing propensity on the part of the examinees can be removed from ability and item parameter estimates. However, in two of the examples above, the contribution of the second examinee parameter to the item characteristic surface was removed by assuming com-

pliance with instructions to answer every item. Accordingly, the item characteristic surfaces were reduced to curves, and, at the same time, the possible effects of differing guessing propensities were eliminated.

Polynomic λ versus Logistic θ

The finite state parameter λ represents ability in a metric that is directly interpretable in a psychological sense, namely, as the amount of knowledge or ability that the examinee has. More specifically, the use of λ is consistent with the remarks of Glaser (1981, p. 935) supporting the use of criterion-referenced testing and expressing "concern for making test scores informative about behavior rather than about *relative performance on poorly specified dimensions*" (italics added). Indeed, the suitability of λ for use in criterion-referenced testing has been addressed in García-Pérez (1989b), where the number of items required for arriving at a mastery decision was determined as a function of examinee response strategy, number of options per item, and test administration format.

The importance of these features of λ (and of finite state polynomic ICCs) can be realized by considering the extent to which θ can be interpreted as a ratio or even interval measure. Lord and Novick (1968, p. 369) point out that "whenever any single item characteristic curve is a monotonic increasing function of θ, it is always possible to transform θ monotonically so that the characteristic curve becomes a normal ogive." The transformed θ is then uninterpretable in any direct psychological sense. Moreover, it is important to realize that this transformation is implicitly and unavoidably made whenever parameters are estimated by fitting data to the normal ogive (or the logistic function). The θs that then result from transformations constrained only to be monotonic can only be interpreted with assurance in an *ordinal* sense.

Additional Benefits of Using Finite State Polynomic ICCs

As shown in the two preceding sections, finite state polynomic ICCs incorporate characteristics of the examinees, the items, and the format of administration of the test in a realistic manner. Also, finite state polynomic models give rise to a measure of ability that is directly interpretable psychologically. But psychological realism and interpretability of λ are not the only practical advantages that can be gained from using finite state polynomic ICCs in place of logistic ones. It is these additional advantages that we consider in this section.

Applicability of IRT Methods to Any Format of Administration of the Test

The mathematical expressions derived using finite state theory are tailored to match any possible specification of the factors represented in Assumptions ii-vi above as they apply to a given test, thus allowing IRT methods to be used with tests administered under any format. One advantage of this fact can be illustrated in the context of the concern with parallel tests on the part of classical test theorists, and the solution provided by IRT to cope with nonparallel test forms. Suppose the same test were administered under two different formats to the

same examinees. Then the two administrations would be strictly parallel (disregarding learning during the test), but two different score distributions would result. The discrepancies between them would result *only* from the differences in the format of administration, since the same examinees and items were involved in both cases. Obviously, these score distributions do not provide direct information about the abilities of the examinees in the group. This is because each format of administration of the test (potentially) gives rise to different and noncomparable response outcomes that are differently related to ability. Under such circumstances, being able to recover the abilities from either of these two score distributions depends on the availability of an appropriate theoretical framework that conveniently accounts for the differences between the administration formats and that prescribes procedures to estimating those abilities from either of the score distributions.

Conventional IRT would apply the same type of ICCs in both cases, which would require changes in either the examinee or the item parameters, despite the fact that the same examinees and items are involved in both cases. Unlike this approach, finite state theory offers the needed theoretical framework and supplies the (different) ICCs that should be used in each case to arrive at the same characterization of examinees and items in terms of their parameters. That scoring methods derived from finite state theory are capable of accomplishing this goal has been confirmed in a dual administration of a test to the same examinees under both conventional and Coombs-type directions (García-Pérez, 1987).

One other advantage of finite state theory as a tool for deriving ICCs is that it yields equations for polychotomous response models as readily as for dichotomous models. This characteristic greatly facilitates the application of IRT to new and varied test administration formats. It also provides a new methodology for the study of the interaction between test format and examinee behaviors with the goal of increased accuracy in the estimation of ability.

Simplified Parameter Estimation

A thorough discussion of parameter estimation for finite state polynomic IRT models would require a separate and lengthy paper. Therefore, in what follows, only major points characterizing these models are presented.

The first of these is that the metric of λ provides unambiguous ability estimates in the case of perfect and zero scores. Unlike what happens with the unbounded θ in logistic ICCs, these scores will result directly in $\lambda=1$ and $\lambda=0$, respectively.

The adaptation of conventional IRT algorithms to the estimation of item and examinee parameters in finite state polynomic models has not as yet been addressed. Nevertheless, this work should not be difficult to carry out, as only a replacement of the mathematical expression to represent the ICC is involved. Moreover, Riefer and Batchelder (1988) have shown how easy it should be to obtain maximum likelihood point estimates and confidence intervals for the parameters of any finite state model.

These concerns aside, very simple methods for the estimation of λ are already available that have proven to yield accurate estimates (see García-Pérez, 1989a; García-Pérez and Frary, 1989). Taking the set of expressions for the probability of every response outcome that arises in a given situation as a system of nonlinear equations, these methods merely involve solving for λ once every probability has been replaced with the empirical proportion of items answered by the examinee in the corresponding response category. In practice, this amounts to finding the single root in the interval [0,1] of what García-Pérez and Frary (1989) called a "scoring polynomial" that is derived from the set of equations under consideration.

Regardless of the approach that is adopted for the estimation of parameters, it is also clear that those procedures will have to be adapted to every particular ICC that finite state theory produces, with special consideration of dichotomous versus polychotomous models. As was pointed out in the last paragraph of the previous section, the properties of the estimates obtained in each case could be taken as a basis for deciding on the optimal administration format for the maximization of accuracy in parameter estimation.

Avoidance of ICCs that Cross

An important side effect of the way polynomic ICCs handle difficulty and discrimination is that any two with the same *c*-intercept will not cross. This may be verified by noting that *p* in Equation 1 is an increasing function of δ and that all of those ICCs are increasing functions of *p*. Hence finite state polynomic ICCs increase monotonically with increasing δ (i.e., the probability of success decreases monotonically with increasing item difficulty). In the case of logistic ICCs, it has been shown analytically by Sijtsma (1988, p. 64) that any two with differing discriminating power *must* cross. This crossing often occurs at extreme values of θ, but, even if the crossing occurs at a θ within, say, [-2,2], it is often the practice to adopt such two- or three-parameter logistic ICCs when they fit the data better than one-parameter logistic ICCs.

In many cases, however, the ubiquitous (so-called "empirical") ICCs that cross may be only the result of applying very powerful curve-fitting techniques to obtain two- or three-parameter logistic functions with differing values of *a*, which, therefore, must cross. In other words, it is the decision to fit the data to a mathematical function permitting the curves to cross that makes estimated ICCs actually cross, sometimes at a θ within [-2,2]. To see how this might happen, suppose that the polynomic ICCs in Figure 3a hold and that responses to the items with δs of .5 and .9 are collected. From the shape of the true ICCs, it is clear that fitting these data to logistic curves will yield much poorer results for the one-parameter function than for the two-parameter function. This is because the true ICCs differ somewhat in slope, which will in turn allow a two-parameter function-fitting algorithm that capitalizes on chance to yield different values of *a* for each item. As a result, their estimated two-parameter ICCs must cross. Hence, artifactually, these items would be considered as evidence that "empirical" ICCs do cross, but in a situation in which the true ICCs do *not* cross.

Apart from this concern, the question of whether to model data with functions that are or are not allowed to cross is theoretical and by no means empirical in nature. We believe that there are strong reasons to prefer as ICCs functions that do not cross. As Wright (1977, p. 103) pointed out in support of the Rasch model, "...we want to think that the probability of success on the harder of two items should always be less than the probability of success on the easier, no matter who attempts the items." This property is ensured only when ICCs do not cross. At the same time, only an item difficulty parameter arising from a framework that yields noncrossing ICCs is well suited to conveying quantitative information about item location in a body of knowledge whose structure can be described by a quasi order, such as that considered by Falmagne and Doignon (1988).

Proper Treatment of Omissions, Guessing, and Partial Knowledge

Conventional ICCs only provide an expression for the probability of getting an item right, which tends to deemphasize the fact that nonright responses can occur in at least two response categories: wrong responses and omissions. As a result, in practice, the treatment of omissions in conventional IRT is limited to categorizing them as either wrong responses or as partially correct responses valued at the inverse of the number of options. Situations exist for which neither treatment would be appropriate. For example, many standardized tests of educational achievement are administered under instructions which indicate that examinees should guess when uncertain regardless of their perceived knowledge. Yet numerous omissions occur, presumably due to lack of examinee motivation or failure to attend to the instructions. To assume that all (or almost all) such omissions reflect total ignorance is certainly questionable. Yet this assumption is required to value omissions at the inverse of the number of options. On the other hand, treatment of omissions as wrong responses would penalize examinees refraining from guessing. Although Lord (1983) proposed a model incorporating a true guessing parameter to account for omissions, available computer programs such as LOGIST (Wingersky, Barton, & Lord, 1982) or PC-BILOG (Mislevy & Bock, 1986) still limit the treatment of omissions to the two choices mentioned above.

Unlike this approach, finite state theory provides an expression for the probability of omitting at the same time that it gives expressions for the probability of getting an item right or wrong, save in cases where omissions do not occur. Hence, finite state theory provides for a proper treatment of omissions under a polychotomous response model. The reason that this is so is that finite state theory models the response behavior appropriately, establishing the contribution of each knowledge state to the probability of each possible response outcome. This can easily be seen by inspection of the right-hand sides of Equations 2-4, where omissions are shown to be the result of failures to guess in cases of total ignorance and partial knowledge. Also, correct and wrong responses resulting from guessing (with various degrees of partial knowledge) occur with the probabilities given by the addends in which γ is a factor, and correct responses resulting from total or partial knowledge have the probabilities given by the remaining addends. Thus, finite state theory allows a distinction to be made between two

different events and their associated probabilities: knowing what the correct answer to an item is, and getting a correct response (by either knowledge or guessing).

Although the distinction between these events and the different cases of partial knowledge underlying them is not usually considered in the use of conventional ICCs, it should be noted that Waller (1989) suggested that the three-parameter logistic ICC can be decomposed into two components which, in turn, can be interpreted as carrying information about the probability of a correct response based on knowledge or as a result of guessing. Under this interpretation, the probability of a correct response with assured knowledge will be provided by a *two*-parameter logistic ICC, while the probability of a correct response from guessing with partial knowledge may be obtained as the difference between this two-parameter ICC and a three-parameter ICC. Finite state polynomic ICCs, instead, model this decomposition explicitly, as specified by the additive terms representing these cases in the functions themselves.

Provision for Independent Tests of Fit

In the introduction, we referred to the suggestion that an ICC should be considered a basic assumption to be tested through goodness-of-fit studies. In this context, it is worth noting that the finite state theory approach to deriving ICCs provides the means for tests of fit in a way that is basically different from the conventional curve-fitting strategy used with logistic ICCs.

Testing the fit of a conventional ICC to data as indicated by a convenient goodness-of-fit statistic raises a fundamental contradiction, since the goodness of the fit is measured *after* parameters have been estimated under the assumption that the model is actually correct. One may wonder to what extent these artifactual estimates force the fit, since it is well known that a good fit can be found even when the source model has nothing to do with the fitted model (see Wood, 1978). Put another way, conventional ICCs cannot really be tested against data, but only fitted to them, since there is no possibility of testing the adequacy of a logistic ICC independently of estimating model parameters. The main reason that this is so is that there are no derivable predictions from logistic ICCs against which empirical data can be contrasted by measuring their agreement with the theoretical expectations.

In contrast, finite state polynomic ICCs can be tested independently. This is true because finite state theory provides expressions for the probability of every possible response outcome to an item. It is this fact that results in testable predictions regarding the relationships among the proportions of responses falling into each response category, predictions that can be tested without recourse to estimating model parameters. Thus, finite state polynomic ICCs have an advantage over conventional ones with respect to Marascuilo's (1988) complaint that models are more often fitted to data than tested against data, since they lead to goodness-of-fit studies and parameter estimation algorithms that are independent of each other.

Further, a major and hard-to-avoid pitfall in the application of conventional IRT to practical problems is the need to discard items that do not fit the assumed model. The fit of logistic

ICCs to a range of empirical data is accomplished by varying the parameters of a single functional expression. In assessing model-data fit, the underlying question is: can reasonable parameters be found such that most of the items pass a goodness-of-fit test? To answer this question, very powerful curve-fitting algorithms are applied in a huge parameter space. Under these circumstances it is not surprising that relatively few items happen to fail the goodness-of-fit test. Nevertheless, even though the goodness-of-fit test nominally pertains to the adequacy of the ICC as an assumption in its own right, it is common practice to discard the item rather than the ICC when the fit is poor. This practice risks producing a test in which items selected on the basis of a statistical criterion may be educationally or psychologically inappropriate (see Goldstein, 1979). This potential outcome occurs as a result of using the same mathematical expression for the ICC of every item in the test, a practice which, in turn, results from being able to employ only a very restricted set of parameters in tailoring conventional ICCs to account for differences among items. Indeed, it simply may not be possible to account for these differences in terms of those parameters.

Finite state polynomic ICCs are not limited in this way. Finite state theory accomodates employing a variety of (valid) assumptions that could permit keeping educationally or psychologically relevant items in a test by properly accounting for their peculiarities. Application of IRT only requires that each item be described by an ICC, with no need for all of them to have the same mathematical form. Although this variation within a test may complicate the estimation procedures, adoption of finite state theory will serve the more critical goal of providing a technique for handling the items that have been chosen to assess the desired educational objectives. By supplying a framework within which the concept of poor-fitting items can be replaced by the more plausible one of inappropriate ICCs, adoption of finite state methods can provide test practitioners with a tool for accomplishing Goldstein's (1979, p. 220) recommended shift of emphasis "towards a development of quantitative assessment techniques which are firmly rooted in qualitative educational objectives."

Discussion

Behind the surface aspects of any psychometric model lies the underlying philosophy of its proponents about how models should be constructed and what should be demanded of them. As for ourselves, what we seek in a model of performance on objective tests is that it be psychologically realistic and directly account for the processes involved in responding to an item (as opposed to those involved in arriving at the response itself). This position motivates the following discussion of the theoretical foundations of finite state polynomic versus logistic ICCs in the context of recently expressed concern about the psychological realism of mathematical models and about the explanatory role of mathematics in psychology.

Perhaps the most appealing feature of polynomic ICCs as compared with their conventional counterparts is that the former are *derived* from an operational definition of knowledge level and a set of realistic assumptions about how test items are constructed and how examinees behave in responding to them, whereas the latter lack this underpinning. At the core of this

difference is the distinction made by Coombs (1983, p. 15) between the descriptive and explanatory role of mathematics in psychology: "The descriptive use of mathematics does not seek to explain an empirical generalization by deducing it from basic (axiomatic) properties of the empirical system. In its explanatory role, mathematics can be used to show that an empirical generalization must necessarily hold." Further advantages of finite state modelling of psychological processes have been discussed in Riefer and Batchelder (1988).

This theory-based approach to developing ICCs is consistent with the recommendations of several authors who have discussed the shortcomings of other approaches. It is clearly in line with Molenaar's (1981, p. 228) request that the role of psychology be dominant over mathematics and statistics in the development of models for achievement testing. Similarly, it conforms with the preference expressed by several authors (e.g., Loftus, 1985; Freedman, 1985, 1987; Marascuilo, 1988) for mathematical models connected to a theoretical framework. rather than simply consisting of a set of equations that data are conveniently found to fit. Indeed, there is more to constructing a psychological model than just getting a good fit.

Extensions to the Finite State Approach

The finite state approach we have dealt with in this paper is a general framework capable of some further improvements in the direction of increased psychological realism or comprehensiveness, three of which we will now outline briefly.

First, as is the usual practice with conventional ICCs, we have not considered the possibility that examinees are misinformed. However, an assumption to this effect would be easy to incorporate realistically into the finite state framework. Toward this end, the options in the item pool would have to be divided into three sets: those whose truth value the examinee knows, those whose truth value the examinee ignores, and those about which the examinee is misinformed. This leads to considering a third examinee parameter, μ, to represent the probability of being misinformed about a given option. In this case, the constraints on these parameters are $0 \leq \mu \leq 1$ and $0 \leq \lambda \leq 1-\mu$. As a result, a third branch would arise from the nodes representing each option in the tree diagrams, the branches now having base probabilities of μ, λ, and $1-\lambda-\mu$. This additional parameter will result in somewhat more complicated parameter-estimation procedures. Nonetheless, as is the case with the fourth parameter in four-parameter logistic functions, μ may often have such a small value that its use in finite state polynomic ICCs may not be worth the trouble.

Second, we have thus far assumed that δ applies to an item and, hence, to each of its options. An alternative view, possibly resulting in increased realism, would be to regard each option as having its own distinct difficulty level. Then each option within an item might have a different value of δ, which in turn would result in different *p*s for each option. This extension of the model would have two different, but related, theoretical implications: it would allow the model to account realistically for the fact that some options are more easily recognized as distractors and, hence, less frequently chosen than others, and it would allow *item option char-*

acteristic curves to be derived for every distractor in an item administered under conventional response directions (as opposed to allowing only the derivation of curves reflecting the probabilities of answering correctly, answering incorrectly or omitting). Thus, as is the case in conventional IRT with the work of Bock (1972) and others, finite state theory is capable of producing IRT models for the nominal response case.

Third, the finite state approach can be extended in the context of speeded tests by the straightforward addition of a speed parameter. An illustration of how this could be accomplished can be found in Link (1982), who used a finite state approach to deriving methods for analysing response times to correct and wrong responses in experiments involving yes-no questions. This framework, either with or without the addition of difficulty parameters, can be directly applied to true-false tests, and it can be further combined with finite state theory to yield models in which both the type of the response and the time it takes the examinee to give it are considered.

A Research Agenda for the Development of Finite State Polynomic IRT models

The main goal of this paper was to present a new kind of ICCs that arise as an extension of finite state theory, a methodology that has already proven useful in modelling performance in objective tests. To make finite state polynomic ICCs usable in practice, the first thing to do is to make procedures available for the estimation of model parameters from responses to objective tests meeting any particular set of conditions. Although, as noted above, this problem has largely been solved with respect to λ (García-Pérez, 1987, García-Pérez and Frary, 1989, 1991), the estimation of δ still has to be addressed.

There are a number of statistics available for measuring the goodness of the fit of a data set to a multinomial model. From the conventional approach to assessing model-data fit in IRT, it has been shown that some of them are more adequate than others (see McKinley & Mills, 1985). As finite state polynomic ICCs allow addressing the issue of model-data fit differently, it remains to be seen which statistics are better for that purpose. Another important line of work will have to do with the empirical comparison of finite state polynomic versus logistic ICCs, not only as to their capability of fitting data but also, and more important, with respect to the predictive validity of the resulting scores.

In addition, and putting together parts of what was mentioned above, use of finite state polynomic ICCs would require the development of a complete model for any situation at hand. Development of a spectrum of models for differing objective testing situations (e.g., with varying numbers of options per item and administered under various formats) would allow a comparison among them to be made with an eye toward maximizing the amount of information about the examinees that is obtained when a set of items is administered.

References

Atkinson, W. H. (1982). A general equation for sensory magnitude. *Perception & Psychophysics*, **31**, 26-40.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29-51.

Coombs, C. H. (1983). *Psychology and mathematics. An essay on theory*. Ann Arbor, MI: University of Michigan Press.

Falmagne, J.-C., & Doignon, J.-P. (1988). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, **41**, 1-23.

Frary, R. B. (1985). Multiple-choice versus free-response: A simulation study. *Journal of Educational Measurement*, **22**, 21-31.

Freedman, D. A. (1985). Statistics and the scientific method. In W.M. Mason and S.E. Fienberg (Eds.) *Cohort analysis in social research: Beyond the identification problem*. New York, NY: Springer-Verlag, pp. 343-390.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics*, **12**, 101-128.

García-Pérez, M. A. (1987). A finite state theory of performance in multiple-choice tests. In E. E. Roskam and R. Suck (Eds.) *Progress in mathematical psychology-I*. Amsterdam: Elsevier, pp. 455-464.

García-Pérez, M. A. (1989a). La corrección del azar en pruebas objetivas: Un enfoque basado en una nueva teoría de estados finitos. *Investigaciones Psicológicas*, **6**, 33-62.

García-Pérez, M. A. (1989b). Item sampling, guessing, partial information, and decision-making in achievement testing. In E. E. Roskam (Ed.) *Mathematical psychology in progress*. Berlin: Springer-Verlag, pp. 249-265.

García-Pérez, M. A. (1990). A comparison of two models of performance in objective tests: Finite states versus continuous distributions. *British Journal of Mathematical and Statistical Psychology*, **43**, 73-91.

García-Pérez, M. A. & Frary, R. B. (1989). Psychometric properties of finite-state scores versus number-correct and formula scores: A simulation study. *Applied Psychological Measurement*, **13**, 403-417.

García-Pérez, M. A., & Frary, R. B. (1991). Testing finite state models of performance in multiple-choice tests using items with "none of the above" as an option. In J.-C. Falmagne & J.-P. Doignon (Eds.), *Mathematical psychology: Current developments*. New York: Springer-Verlag.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, **36**, 923-936.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, **5**, 211-220.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hanna, G. S. (1975). Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, **12**, 175-178.

Hutchinson, T. P. (1977). On the relevance of signal detection theory to the correction for guessing. *Contemporary Educational Psychology*, **2**, 50-54.

Link, S. W. (1982). Correcting response measures for guessing and partial information. *Psychological Bulletin*, **92**, 469-486.

Loftus, G. (1985). Johannes Kepler's computer simulation of the universe: Some remarks about theory in psychology. *Behavior Research Methods, Instruments, & Computers*, **17**, 149-156.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1983) Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, **48**, 477-482.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Marascuilo, L. A. (1988). Introductions to model building and rank tests. *Contemporary Psychology*, **33**, 794-795.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, **6**, 379-396.

McKinley, R. L., & Mills, C. N. (1985). A comparison of goodness-of-fit statistics. *Applied Psychological Measurement*, **9**, 49-57.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, **42**, 725-737.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Molenaar, I. W. (1981). On Wilcox's latent structure model for guessing. *British Journal of Mathematical and Statistical Psychology*, **34**, 224-228.

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, **47**, 355-366.

Mosier, C. I. (1941). Psychophysics and mental test theory. II. The constant process. *Psychological Review*, **48**, 235-249.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, **95**, 318-339.

Samejima, F. (1981). *Final report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item* (ONR Contract No. N00014-77-C-0360). Knoxville, TN: University of Tennessee.

Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory*. Doctoral dissertation. Amsterdam: Free University Press.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, **51**, 567-577.

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement*, **13**, 233-243.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST 5.0. Version 1.0. User's Guide*. Princeton, NJ: Educational Testing Service.

Wood, R. (1978). Fitting the Rasch model - A heady tale. *British Journal of Mathematical and Statistical Psychology*, **31**, 27-32.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, **14**, 97-116.